DOCUMENT RESUME

ED 467 846 TM 034 313

TITLE	Setting Cut Scores on Large-Scale Assessments. The State Board Connection Issues in Brief.
INSTITUTION	National Association of State Boards of Education, Alexandria, VA.
SPONS AGENCY	CTB / McGraw-Hill, Monterey, CA.
PUB DATE	1999-07-00
NOTE	14p.; Vol. 19, No. 2.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	*Cutting Scores; *Educational Testing; *Standard Setting; *State Programs; Statistical Analysis; Test Construction; *Test Use; Testing Programs
IDENTIFIERS	*National Association of State Boards of Education

ABSTRACT

This Brief builds on the work of the National Association of State Boards of Education 1997 Study Group on State Assessment Systems by examining one of the state board actions that is most likely to capture the public's attention: setting cut scores on state assessments. This process involves a large measure of human judgment and politics and a sense of what the system will bear, as well as statistical principles. There are methods and practices that policymakers should understand in order to have as valid and reliable a system as possible. Test-centered, examinee-centered, and compromise methods are reviewed. It must be recognized that all methods of setting cut scores are technically imperfect, and that standards need to be clear. Choosing the most effective method of setting cut scores depends on the context of state standards and the format of state tests. (Contains 11 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

NASBE
NATIONAL ASSOCIATION OF STATE BOARDS OF EDUCATION

The State Board Connection

ISSUES IN BRIEF

July 1999
Vol. 19, No. 2

ED 467 846

This *Issue Brief* builds on the work of NASBE's 1997 Study Group on State Assessment Systems by examining one of the state board actions that is most likely to capture the public's attention: setting cut scores on state assessments. This process—which essentially answers the question of "how good is good enough?"—involves a large measure of human judgement, politics, and a sense of what the system will bear, as well as statistical principles. Yet as the *Issue Brief* explains, there are methods and practices that policymakers should understand in order to have as valid and reliable a system as possible.

NASBE would like to thank CTB/McGraw-Hill for their support in the development and publishing of this *Brief*.

Setting Cut Scores on Large-Scale Assessments

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Welburn

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034313

BEST COPY AVAILABLE

NASBE
NATIONAL ASSOCIATION OF STATE BOARDS OF EDUCATION

The State Board Connection
ISSUES IN BRIEF

July 1999
Vol. 19, No. 2

This *Issue Brief* builds on the work of NASBE's 1997 Study Group on State Assessment Systems by examining one of the state board actions that is most likely to capture the public's attention: setting cut scores on state assessments. This process—which essentially answers the question of "how good is good enough?"—involves a large measure of human judgement, politics, and a sense of what the system will bear, as well as statistical principles. Yet as the *Issue Brief* explains, there are methods and practices that policymakers should understand in order to have as valid and reliable a system as possible.

NASBE would like to thank CTB/McGraw-Hill for their support in the development and publishing of this *Brief*.

Setting Cut Scores on Large-Scale Assessments

Introduction: The Expanding Field of Assessment Standards

The standards- and assessment-driven education reforms underway in states across the country have taken hold more quickly, and spurred educational change more rapidly, than perhaps any other educational reform of the past century (Bond, 1995). Although standardized tests have been used by schools and districts to evaluate and sort students for almost a century, what is new about today's assessment programs is their sophistication, variety, and emphasis on standards.

Standards were previously defined implicitly by passing scores on tests. Today, however, theyare custom-designed to reflect clearly articulated public expectations of schools, students and teachers. At the same time, they are aligned with assessments that measure the *specific* standards, often through a diverse set of questions that includes not only multiple-choice but also new types of essay, short-answer and performance tasks that can measure knowledge application and skills development. Truly, schools and entire state education systems today are being reformed in large part by standards- and assessment-driven initiatives.

> Cut scores...define at least two important components of a standards-based system—they provide a clear measure of "how good is good enough" and provide a standard yardstick by which to measure progress.

If standards are the skeleton around which states build their education system and assessments are the muscles that bring the system to life, then *cut scores* are the vital signs by which the quality of life is evaluated. Cut scores, the actual numerical limits applied to student performance on an assessment, define how well students, teachers, schools, or the education system are performing, tie student performance on an assessment directly to the standards, and fundamentally define at least two important components of a standards-based system. In

Types of Assessment

Norm-referenced tests (NRTs) compare individual performance against the performance of a representative national sample. Scores from NRTs are usually reported according to "percentile" rank (a student's percentile rank on a test is the percent of students he or she outscores). This means that students are evaluated in relation to each other rather than to a "standard." As a result, cut scores on NRTs cannot measure student performance according to content standards.

Criterion-referenced tests (CRTs) compare student performance to clearly defined curricular objectives—standards. Assessment results are usually reported as a pre-defined level of performance (such as "advanced" or "novice") or a numerical score. CRTs require the development of meaningful learning objectives that are keyed to specific assessment items, and CRTs therefore assess what students know and can do rather than how students compare with their peers. Cut scores on CRTs are usually carefully calibrated to content and performance standards.

Multiple-choice assessments require students to select their responses from among a set of specific, predetermined choices. When carefully designed, multiple-choice assessments can provide reliable information about what students know. Usually, multiple-choice assessments are scored by computers, providing quick, cost-efficient, and accurate results. However, their characteristic computer scoring does not mean that multiple-choice assessments are "value-free," "neutral," or "unbiased"; on the contrary, setting cut scores on multiple-choice assessments is still a human endeavor and therefore still open to human judgement.

Performance assessments require students to formulate an original response to a question and to communicate that response through the performance of some act. For example, a performance assessment may ask a student to produce a written essay or short answer, model, diagram, or persuasive speech. Performance assessments are usually criterion-referenced. Standards that emphasize application of knowledge and expression of ideas are most effectively evaluated with performance assessments. However, assessment experts debate the overall reliability of various types of performance assessments, in part because of the difficulty in setting cut scores, creating standardized scoring protocols, and ensuring consistency among scorers.

Because no one test can "do it all," states should consider "implementing an assessment system with more than one test that balances performance assessment items with more traditional multiple choice" (NASBE, 1997).

other words, they provide a clear measure of "how good is good enough" and provide a standard yardstick by which to measure progress.

States are learning that in large part the extent to which assessments are aligned with standards, both in terms of their content and in terms of the performance levels implied by their cut scores, determines whether the standards actually become the central goals that drive teaching and learning in the state. Since consequences are usually distributed according to specific assessment results rather than to the standards themselves, in systems where the assessment system, cut scores, and standards are not aligned, teaching and learning are usually more affected by the content of the assessment and the cut scores associated

with it than the standards themselves. As a result, it is important for states to build assessments and develop cut scores that support their standards rather than undermine them.

Fundamentally, there are several key ways in which good cut scores support the success of state standards and assessment systems:

- **Cut scores provide a system to measure existing levels of performance.** An assessment system cannot measure performance according to standards without applying a specific system of measurement that uniquely calibrates the two and says what level of test performance meets or exceeds the standards. This system of measurement is defined by a set of cut scores.

- **Cut scores provide a yardstick by which policymakers can set future achievement goals.** Although setting standards that describe what students should know and be able to do is one step in creating a standards-based system, until those descriptions are translated into actual quantifiable measures policymakers cannot discern where student performance lies in relation to standards or where they would like it to progress over the next three, five or ten years. Cut scores allow policymakers to set measurable goals for improvement across the education system.

- **Cut scores, and the scoring rubrics associated with them, provide students and teachers with actual examples of the kind of work expected of them under the new standards.** Cut scores provide clear, concise, quantifiable definitions for content and performance standards that define for teachers and students, in terms directly related to actual student work, how good is good enough to meet the standards. This definition is an important way to let teachers and students know exactly what is meant by standards.

- **Cut scores provide a yardstick by which assessment results can be understood.** Particularly with new criterion-based performance standards and corresponding performance assessments that often report student performance in terms of hierarchical grouping such as "advanced" or "novice," it can be difficult for parents and members of the public to interpret exactly how well students are doing. Clearly demarcating cut scores that define particular levels of performance can help the general public make sense of new and increasingly complex standards and assessment criteria.

As assessments have become important tools in high-stakes decisions concerning school reform and improvement, teacher evaluation, student evaluation, and promotion, the fairness of particular cut scores as well as the appropriateness of the methods used to derive them have become prominent issues of debate among measurement experts, educators, policymakers, and the general public. This issue brief examines issues that face state policymakers as they establish cut scores for their state assessment programs that: 1) support state standards; 2) are valid and reliable; 3) are fair and equitable; and 4) are likely to have public support.

Selecting a Method for Setting Cut Scores

Although for decades cut scores on tests have been applied to individuals for high-stakes decisions such as graduation, course completion, or college placement, there was until recently little attention paid to *how* passing scores on particular tests were established—in other words, what "standard" the cut scores actually represented and how that standard was derived. In recent years, as public discussion about standards has taken hold, so too has public discussion about cut scores, realization of their politically charged nature, and desire to understand the rationale behind the methodology of setting cut scores. In essence, when policymakers "set the bar" for rewards and sanctions in terms of cut scores in a standards-based system of assessment, those affected by the level of the bar often demonstrate a keen interest in knowing why the bar is set precisely where it is. To further complicate matters, rewards and sanctions, particularly at the school level, can be doled out in relation to whether scores are improving or not, rather than whether scores reach an absolute bar. Setting up innovative and unfamiliar ways of evaluating achievement has been known to generate high levels of public interest in understanding cut scores and their consequences.

Currently, there are at least 50 methods that test developers and measurement experts use to set assessment cut scores (Berk, 1995). This variety arises due to the fact that, although testing experts may prefer certain methods and researchers may point out ways in which methods differ, no single method is universally "best" or "most accurate." In fact, it is likely that no single method will ever surface as universally "best" because determining acceptable performance levels and translating them into cut scores on an assessment *always* involves some decisions that are arbitrary and others that are purely based upon subjective human judgement (Ebel, 1972). Furthermore, there is no "gold standard" to which the results of different cut score methods can be compared, so although different cut score methods often give rise to different cut scores, it is impossible to tell which cut score is actually closest to the "real" standard. Even when set by reasoned, well-informed judges, cut score decisions are always the result of human interpretation and opinion; as a result, establishing and evaluating cut scores is always, to some extent, relative.

> ...no single method [of determining assessment cut scores] is universally "best" or "most accurate."

The methods currently in use to set cut scores for state assessment systems can be described in one of three ways: test-centered, examinee-centered, or combined.

- ***Test-centered methods*** of setting cut-score standards require judges to evaluate how difficult each item on an assessment will be for minimally competent examinees. In other words, judges examine test items, not test answers, and evaluate the items on criteria that are based upon the standards set by the state board. After each judge decides how minimally competent examinees should perform on each test item, the decisions of all judges across every test item are combined to arrive at cut scores that define different performance levels.

- ***Examinee-centered methods*** are those based upon judgements of individuals' actual performances on an assessment. In examinee-centered methods, judges evaluate individuals' actual test performances according to established performance standards and identify the point at which performances are "borderline," that is, the point at which examples of student work begin to clearly transition from one performance level to another. The median or average score given by judges to individuals who perform at the transition between two performance levels becomes a cut score.

- ***Compromise methods*** employ both absolute and normative standards to set cut scores. Often, there are implicit expectations not only about the content knowledge that individuals should be able to demonstrate, but also about the percent of individuals who should meet a standard. Compromise methods take both of these measures, the absolute content standards, and the normative expectations of passing rates, into account when figuring cut scores.

Until recently, most assessment cut scores were set using test-centered methods, in large part because test-centered methods are particularly amenable to multiple-choice tests; as a matter of fact, some test-centered methods can only be applied to multiple-choice questions. Only recently, particularly as a result of the growth of new testing methods that include *performance* evaluations, have examinee-centered methods come to be used as frequently as test-centered methods. Many measurement and evaluation experts contend that examinee-centered methods are more able to evaluate the variety of likely answers on performance assessments and take account of differential point values assigned to different test questions than are test-centered methods.

In general, assessments that elicit responses from examinees that are rich and complex or that require performances that are difficult to break down into small component tasks are usually most effectively complemented by examinee-centered methods of setting cut scores. This is because it is difficult for judges, without the example of actual student work in front

of them, to envision all of the complex ways in which specific skills can be integrated to produce a single complex product (Kane, 1995). As performance assessments have grown in popularity, in part because of their perceived positive impact on teaching and learning methods and in part because of perceived advantages in terms of validity, examinee-centered methods of standard-setting have grown in popularity. Still, some states like Vermont and Kentucky found that the cut scores they set through examinee-centered methods, and the scoring rubrics and evaluation protocols associated with implementing the cut score criteria, did not lead to evaluation results that were consistent enough to be used for some types of high-stakes decisionmaking.

Assessments that break examinee performances into small parts and base scoring on a large number of small and discrete tasks are largely amenable to test-centered methods of setting cut scores. This type of test is one in which writing ability, for example, would be evaluated by several questions concerning punctuation, spelling, and syntax rather than the production of a written paragraph or essay. Most test-centered methods were developed to be used with exactly these kinds of tests. Using these methods on performance tests, however, often wipes out one of the major advantages of performance tests—the ability of judges to make "holistic judgements about performances" (Julian, 1993). In other words, the reasons that test-centered methods are easy to use on objective tests are the same reasons they are difficult to apply meaningfully to performance tests (Kane, 1995). On the other hand, at least one promising test-centered method has been created specifically for use with performance tests and is being used in Kansas (Mehrens, 1995).

Regardless of which method is utilized to set cut scores, it is important to recognize that every method has strengths and weaknesses. In fact, it is difficult to draw concrete conclusions about which method of setting standards is uniformly better or worse. In general, the method of setting standards depends upon the type of test and its intended use, so state policymakers need to ensure that the method of setting cut scores complements the type of skills and knowledge that are defined by state content and performance standards and evaluated on the state assessment. Furthermore, savvy state policymakers understand that whichever method is employed, there is an almost certain guarantee that the results of that method will differ to some extent if compared to the results of other methods, both in terms of the level of cut score they set and the level of consistency they get among judges' evaluations of test items or responses. As a result, it is important for policymakers to concentrate not only on the way in which cut scores are set, but once they have been set, to consider those cut scores in terms of their defensibility, primarily in terms of their *validity, reliability, fairness,* and *political acceptability.*

Issues to Consider

- ***All methods of setting cut scores are technically imperfect.*** At the moment, examinee-centered methods seem to have more technical challenges, in large part because these methods are newer and have been studied less (Kane, 1995). One could also argue, however, that considering the extent to which test-centered methods have been evaluated and refined over decades of use, they should have considerably fewer technical problems than they do, and so with a few years of concerted study of examinee-centered methods, the technical gaps between the two types of cut score-setting procedures could close. Either way, if is important for policymakers to recognize that there is no "silver bullet" in terms of setting cut scores, and all cut scores rely on some level of subjectivity.

- ***Standards need to be clear.*** The quality of cut scores can only be as good as the standards upon which they are based. In other words, it is important to make sure that content standards are easily translatable into performance standards, and performance standards make clear to the judges exactly what should be expected in terms of assessment performance at every performance level. Yet many states have developed standards that are neither sufficiently rigorous nor specific enough to direct assessment development or the setting of cut scores. Unless the system lines up behind clear standards, it is likely to be the assessments, more than the

standards, that shape the behavior of students and teachers.

- ***Choosing the most effective cut score method depends upon the content of state standards and the format of state tests.*** While no method of setting cut scores is uniformly superior to all other methods, there are certain methods that are more appropriate to use with particular types of test questions. Understanding general approaches to setting cut scores, and the circumstances under which each approach is preferred, can help policymakers ensure that the cut scores established for their state assessments have been set appropriately.

Setting Cut Scores: A Lesson from Michigan

In Michigan, the widespread refusal of parents to let their children take a high school proficiency test threw the establishment of a statewide exam into a tailspin. In 1997, many parents boycotted the exam because they believed it offered no rewards for college-bound children. The assessment, given to juniors, was created to measure high standards and was intended to mollify critics who asserted that high school graduates lacked basic skills. On each section of the criterion-referenced test, students scored *proficient*, *novice*, or *not-yet-novice*, and the results went on their school transcripts. Students who scored proficient received gold seals on their diplomas.

Although the tests were not required for graduation, parents were concerned that including the test results on permanent transcripts could penalize children who had done well in their high school courses, scored well on the SAT, and were heading to competitive colleges and universities. Preliminary test results seemed to confirm their fears: in 1996, only 34 percent of students scored proficient in writing, 47 percent in math, 32 percent in science, and 40 percent in reading. Stories of students who scored "novice" on one or more components of the Michigan test, but who scored exceptionally well on the ACT and SAT and were accepted to highly prestigious universities, fueled parental concerns about the validity of the test.

In response to parental concerns, the legislature and state board of education made several changes to the way the test was administered, including: establishing new performance levels ("*exceeded state standards*," "*met state standards*,""*endorsed at basic level*," or not endorsed); removing any endorsement from the diploma; establishing a review and appeals process for parents of students who receive low test scores; shortening the duration of the tests; moving the administration of the test to the end of the 11th grade so students have more opportunity to learn the curriculum; and expanding a Parent Report to help parents understand their child's test result.

While the Michigan experience has been viewed as a lesson in the importance of public engagement, it can also be viewed as a cut score issue, since it was the cut scores and how the performance levels were defined, more than the standards or the test content, that ultimately placed what many thought was an inordinate number of students into the "novice" category.

Reliability and Validity of Standard-Setting

Regardless of the specific method used to set cut scores, every method needs to be checked for *validity* and *reliability*. In general, validity measures ensure that cut scores really represent the intended standards; reliability measures, on the other hand, establish that judges' decisions about where to set cut scores are consistent and replicable. Both measures are indispensable in building a system of standards-based assessment that is fair and credible.

Checking cut score-setting methods for validity and reliability is not the same thing as establishing validity and reliability for the assessment itself, and so a discrete set of measures and observations—in addition to those done on test items—needs to be completed in order to ensure that cut scores are valid and reliable. In general, there are at least four different measures that policymakers should look for in order to ensure that the cut scores they apply to individuals taking an assessment are valid and reliable.

In terms of reliability, at least two important measures assess the extent to which judges make decisions about cut scores that are consistent and replicable (Berk, 1995):

- **Intrajudge Reliability**—Intrajudge reliability is an important measure of the extent to which individual judges consistently apply the same criteria to their decisions about test items or work samples. The extensive training judges receive to match items with performance levels, combined with feedback or discussion during the cut score process, is meant to ensure the internal consistency of individuals' judgements across a test or set of work samples. Measurement experts are usually very concerned about ensuring intrajudge reliability and will monitor it with reliability checks across the standard-setting process.

- **Interjudge Reliability**—Measures of interjudge reliability are concerned with the consistency of opinion across a judging panel. Ideally, judges' decisions about test items or work samples converge. But convergence can be disrupted by several factors, including providing judges with standards or achievement levels that are ambiguously defined, selecting judges poorly, or training judges inadequately. Figuring out the extent to which variations between judges' evaluations affect the replicability of judges' decisions is an important way to evaluate the success of the standard-setting process.

Cut scores also need to be evaluated in terms of their validity. Validity measures the extent to which cut scores truly represent the intended standard—in other words, whether the cut scores measure what they are intended to measure. Establishing validity of cut scores is problematic, in large part because of the fact that there is no concrete criteria that can be used to evaluate the validity of a cut score. Still, there are at least two types of validity that measurement experts often try to establish in relation to cut scores (Berk, 1995):

- **Consequential Validity**—Consequential validity gauges the consequences of judges' cut score decisions. After judges establish cut scores, they are often adjusted up or down based upon the economic, political, social, and educational consequences of applying the cut scores that were set. By using pilot test results, policymakers can get some idea of how students would be sorted according to the proposed cut scores and—if the consequences of applying the cut scores are untenable politically, socially, or economically—the cut scores may be adjusted. When evaluating the consequential validity of cut scores, policymakers may consider a variety of issues such as any adverse impact the cut scores may have on particular groups of students; the impact that applying the cut scores may have on school initiatives, school morale, or staff stability; the impact that reporting results based on the cut scores may have on public credibility of the standards and assessment system; or ways in which implementing the cut scores may affect the economic opportunity of large numbers of students or certain communities.

Among some testing experts, consequential validity is controversial; while all experts agree that

the consequences of setting particular cut scores are important considerations in the decision-making process, experts disagree whether this rather pragmatic and often political issue should be considered a "validity" issue. For state board members, it is important that the consequences of cut scores be studied and seriously considered, regardless of the terminology ascribed to the actual process.

...Assessment systems that fail to establish good reliability and validity of their cut scores can fall into trouble rapidly.

- **Evidential Validity**—Evidential validity refers to the extent to which students' performance on an assessment actually represents the extent of their knowledge in the content areas being tested. Cross-referencing student performance on an assessment with other measures of knowledge such as teacher expectations, standardized test scores, or course grades can give policymakers an idea of whether cut scores will derive test results that correspond with other publicly accepted measures of student performance. If cut scores often disagree with results that would be expected according to other measures of student performance, then the cut scores may be said to have low evidential validity.

One of the most contentious issues concerning the validity of tests and cut scores is *fairness*. Fairness is often discussed in terms of *bias* and *adverse impact*. A test is said to be biased if a particular group of students performs significantly lower than other students on a test, and the low performance can be attributed to some aspect of the test itself. In terms of setting cut scores, tests can be made biased by inappropriate cut score decisions by unrepresentative judges, failure to take account of ways in which dialects and culture may affect test answers, or assignment of rigorous requirements and significant point values to skills that are not supposed to be tested on a particular assessment, such as requiring extensive essay writing on a mathematics test. Many issues of "fairness" are also issues of validity, since they involve creating a test or setting cut scores that measure something other than what they are supposed to measure.

When an assessment is biased to the extent that it has an *adverse impact* on one or more groups of students (meaning that it results in the disproportionate denial of some reward or honor to students in the impacted group) the test is often considered unfair. Because every state assessment needs to be as fair as possible, it is important for state policymakers to collect and analyze data about bias and adverse impact throughout the assessment process, from the creation of test items to the setting of cut scores to the administration and scoring of test results. Suggestions for minimizing bias through extensive public input and a rigorous feedback process are included in the following section, "Politics and Public Perceptions in Setting Cut Scores."

Establishing validity and reliability of the cut score method is a necessary component of building a standards and assessment system that is fair, credible, and legally defensible. Conversely, assessment systems that fail to establish good reliability and validity of their cut scores can fall into trouble rapidly. For example, in Arizona a statewide performance assessment was suspended in 1995 in large part because initial results from the performance assessment did not correlate with the results of another more traditional part of the assessment program; in other words, the results lacked evidential validity. Fine-tuning the cut score process to more concretely establish evidential validity before administering the performance assessment may have made it less vulnerable to criticism and more likely to withstand widespread public scrutiny.

Effective state policymakers understand that validity and reliability issues span the range of steps involved in creating an assessment system, including setting cut scores. By closely monitoring the cut score process to ensure that highly trained, representative judges consistently derive cut scores that are likely to correspond with the results of other measures of student performance and not lead to untenable adverse consequences, state policymakers increase the chances that their assessment system will be fair, credible, defensible, and able to withstand public scrutiny.

Issues to Consider

- **Standard setting, based on current knowledge, will always lead to errors in classification.** These errors occur when students fail to reach a cut score although they actually possess the skills and knowledge required by a particular performance level, or when students pass a cut score without actually possessing the requisite knowledge for the corresponding performance level. These types of errors occur in any assessment administration for several reasons, including scorer error, judge error, student guessing, and student error. Any state assessment system needs to balance the need to minimize errors in classification (which can vary according to type of test, the way in which cut scores were set, or the actual scoring method employed) against the need to set clear cut scores and make them stick. Establishing the validity and reliability of assessment cut scores is an important way to minimize the number of errors in classification that result from an assessment.

- **Judges need to be very carefully selected and trained.** Selecting the most credible judges possible and carefully training them to ensure that, as much as possible, they share the same understanding of what standards they are looking for is one of the most important components of setting cut scores. Many states have found that credible judges include a wide array of representatives, including teachers, curriculum specialists, parents, university faculty, and general public and business representatives. Furthermore, those involved in selecting and training judges need to keep in mind that training effectiveness and training quantity often determine the effectiveness of the entire process of setting standards (Berk, 1995). State policymakers who pay attention to the make-up and preparation of cut score judging panels find they can more easily ensure that the cut scores are valid, reliable, and generally perceived as fair.

- **All methods of setting standards depend upon judgements.** In every method of setting cut scores, judges rate test items or test answers according to the content and performance standards usually established by the state board of education. Translating descriptions of student performance into numerical cut scores is always laden with values, assumptions, and ultimately imperfect human judgement. Policymakers need to be aware that no cut score represents a true standard.

Politics and Public Perceptions in Setting Cut Scores

State efforts at establishing standards and developing related assessment systems have occurred partly as a response to increasing public skepticism about the ability of large government bureaucracies to improve education. Yet ironically, creating standards, developing assessments, and setting cut scores are in themselves major governmental undertakings that have in some cases only reinforced the public's skepticism of the ability of the state education agency to reform itself. While states are keen to prove that the new standards-based systems are truly different—that they actually free schools, teachers, and students from prescriptive processes and burdensome regulations in favor of a state role that concentrates on outcomes—public buy-in has sometimes been slow and public scrutiny fierce. As a result, the process of setting cut scores is often keenly observed and highly political.

The political nature of building assessments and setting cut scores is exacerbated by the fact that experts who set cut scores and relate them to standards concede that the process is not an exact science; as described in the previous section, different methods of setting scores and relating them to standards give rise to different cut scores, and every method of setting cut scores is affected to some extent by hu-

> ...setting cut scores is both subjective and scientific and is likely to be as political and ideological an undertaking as setting standards themselves.

Cut Scores and Massachusetts Assessments

When Massachusetts released results from its new Comprehensive Assessment System, policymakers hoped that months of public preparation would pay off. Like many states that are administering tough new tests, policymakers have been investing considerable time and energy in getting people to understand what the new tests are and what they are not. One thing that the Massachusetts tests are is difficult: in the first administration of the assessment, 50 percent of 10th-graders failed the math portion of the test, 40 percent of 8th-graders failed the science test, and two-thirds of 4th-graders were categorized as "needs improvement."

Massachusetts is an example of a state where officials decided to risk public backlash against their testing program for the sake of preserving a set of cut scores that they believed in—even though policymakers conceded that the cut scores established exceptionally high expectations for students.

So far, the public in Massachusetts appears not to have launched the type of backlash that has stymied challenging statewide assessments in other states, a fact that may point out that the success of standards and assessment systems may rest in part on how they are presented to the public. Massachusetts undertook a massive public awareness campaign prior to releasing the test results, including making test questions available to the public to highlight how tough the tests are, releasing an informational video and brochure that explained the test before it was administered, including test study guides on the education department's web site, and orchestrating public statements from the governor, state school superintendent, state board, and others that braced the public for low results.

man judgement. As a result, setting cut scores is both subjective and scientific and is likely to be as political and ideological an undertaking as setting standards themselves.

Many states have found that the best laid assessment plans have been undone by a political outcry of one sort or another, be it a result of the type of test administered, the content of state standards, the setting of cut scores, the groups in which students are to be categorized, the consequences attached to the test, or the test's perceived fairness. The good news, however, is that the experiences of several pioneer states have generated useful guidelines that states can use to help minimize the negative politics and increase public support as they implement these reforms.

1. Broad public engagement is necessary across the process of establishing a standards-based system.

Because the types of assessments, the way scores are reported, and the consequences attached to assessment results may be new to the public—and because applying cut scores to administer consequences to students, schools, or the system may be politically controversial and open to accusations of bias—broad engagement of key constituents at every stage of the process is imperative in order for the resulting system to be stable and credible. Constituents that may participate at various points in the standards/assessments/cut scores processes include: teachers, parents, students, school administrators, local school board members, higher education faculty and administration, education researchers, state legislators, the governor, business representatives, and the media. Involving a diverse set of constituents as early and as long as possible is one way to guard against unfair decisions across the process.

2. The process of developing tests and setting cut scores needs to be open to public scrutiny.

Secrecy, justified in part by the need to retain the integrity of an assessment, has long been a hallmark of the assessment community (McDonnell, 1997). But given the high stakes, new methods, and higher standards included in many state assessment

systems, the public has become understandably reluctant to accept assessment systems, including cut scores and their consequences, on blind faith. An open process of setting standards, developing assessments, and defining cut scores can ensure that all sides have a voice and can help ensure that assessment systems are based upon consensus. Furthermore, opening the process to public scrutiny and comment may be an effective way to pinpoint potential causes of bias and adverse impact before the assessment system is first administered.

3. There is no substitute for taking your time.

Because of the complexity and inventiveness of many standards-based assessment systems, not to mention the systems' emphasis on spurring long-term systemic change, establishing the system and fine-tuning it is a long-term process. Nonetheless, at least in part because of the strong political dimension that often requires political expediency to be placed above technical advisability, some states have been forced to implement standards-based systems rapidly. In California and Kentucky, for example, "the assessment designers had to balance daunting technical challenges against the political reality that elected officials expected to see performance assessments on-line within a very short time frame.... Given the constraints... considerably more time was needed for test development" (McDonnell, 1997).

In terms of setting cut scores, taking your time means making sure that content and performance standards are sufficiently clear and precise to inform the development of test items and define cut scores; developing and evaluating the ways in which judges are selected and trained before they are allowed to set cut scores; judging the proposed cut scores in a variety of ways for validity and reliability; preparing the public for the consequences of implementing particular cut score decisions; and evaluating the intended and unintended outcomes of implementing cut scores with an eye toward fine-tuning the system.

4. Every system will have its critics.

Effective policymaking requires boards to pay attention to the concerns of all groups in the state. But effective policymaking also requires boards to concede that even the best plan is likely to meet with disapproval from several small but potentially vocal groups. Policymakers need to devise effective, proactive ways to engage the issues of their critics and ensure that all voices in the debate are heard before decisions are made. In addition, many states have turned criticism to their advantage by mounting concerted communications efforts that help to diffuse criticism by explaining the system in positive terms and concentrating on reaching those whose minds are not yet made up, rather than continuing to address the concerns of those who will never be convinced.

5. Strong political leadership is absolutely necessary.

Establishing a system of standards and assessment—and setting cut scores that will take hold and become accepted measures of the content and performance standards—takes a commitment that may span several political administrations. As such, there needs to be a cadre of political leaders who exhibit the courage to "stay the course" and communicate to the public the idea that systemic change takes time and requires flexibility and tolerance from everyone who hopes to benefit from the new system down the road. Besides supporting the long-term nature of standards- and assessment-based reform themselves and encouraging everyone to view cut scores as representing goals and yardsticks rather than as absolute measures of success or failure, state board members can also support politicians, department personnel, local board members, and educators who encourage a long-term commitment to the system.

An open process of setting standards, developing assessments, and defining cut scores can ensure that all sides have a voice and can help ensure that assessment systems are based upon consensus.

Issues to Consider

- **Equity and fairness need to be established.** Public rejection of state assessments is often caused by a perception that the test is unfair or biased. Putting public fears of adverse impact to rest—and convincing the public that high standards as represented by the test are in everyone's interest—are important components of establishing long-term stability for the testing system.

- **There is not one right way to set standards, and so setting standards is likely never to be without controversy.** The fact is that every standard-setting method will produce errors, and so there will probably always be individuals who feel that a particular method is unfair. State policymakers need to recognize this inherent weakness in setting standards, ensure that protections have been put in place to minimize errors as much as possible, and continuously collect data to refine and improve the establishment and implementation of standards.

- **Allow time for standards to become established before tinkering with the standard-setting procedure.** Every procedure is likely to yield results different from others. Recognizing this, it is wise for policymakers to stick with one procedure for some time; otherwise, standards start to fluctuate so much that they lose credibility. Standards are never absolute. This makes setting standards a process rather than a single act, and that process needs to occur slowly, so that consistent understandings about standards can be accepted by the public before revisions are introduced.

References

Berk, R. (1995). Standard setting: The next generation. In *Proceedings of the joint conference on standard setting for large-scale assessments.* Washington, DC: National Assessment Governing Board/National Center for Education Statistics.

Bond, L. (1995). Ensuring fairness in the setting of performance standards. In *Proceedings of the joint conference on standard setting for large-scale assessments.* Washington, DC: National Assessment Governing Board/National Center for Education Statistics.

Claycomb, C. & Kysilko, D. (1997). *Issues and trends in statewide assessments.* Alexandria, VA: National Association of State Boards of Education.

Ebel, R. (1972). *Essentials of education measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Johnston, R. (1997). Just saying no. *Education Week,* April 9.

Julian, E. (1993). *Standard-setting on performance examinations.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of the joint conference on standard setting for large-scale assessments.* Washington, DC: National Assessment Governing Board/National Center for Education Statistics.

McDonnell, L. (1997). *The politics of state testing: Implementing new student assessments.* Los Angeles: National Center for Research on Evaluation, Standards, and Testing.

Mehrens, W. (1995). Methodological issues in standard setting for educational exams. In *Proceedings of the joint conference on standard setting for large-scale assessments.* Washington, DC: National Assessment Governing Board/National Center for Education Statistics.

National Association of State Boards of Education. (1997). *The full measure: Report of the NASBE study group on statewide assessments.* Alexandria, VA: Author.

White, K. (1998). New tests send bay state schools scrambling. *Education Week,* April 8.

NASBE
NATIONAL ASSOCIATION OF STATE BOARDS OF EDUCATION

Issues in Brief are developed and produced at the National Association of State Boards of Education, 277 South Washington Street, Suite 100, Alexandria, VA 22314 (703) 684-4000. Brenda Welburn, Executive Director; David Kysilko, Editor; Carla Claycomb, Author.

All rights reserved.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").